

ATTACHMENT B

Exhibit A

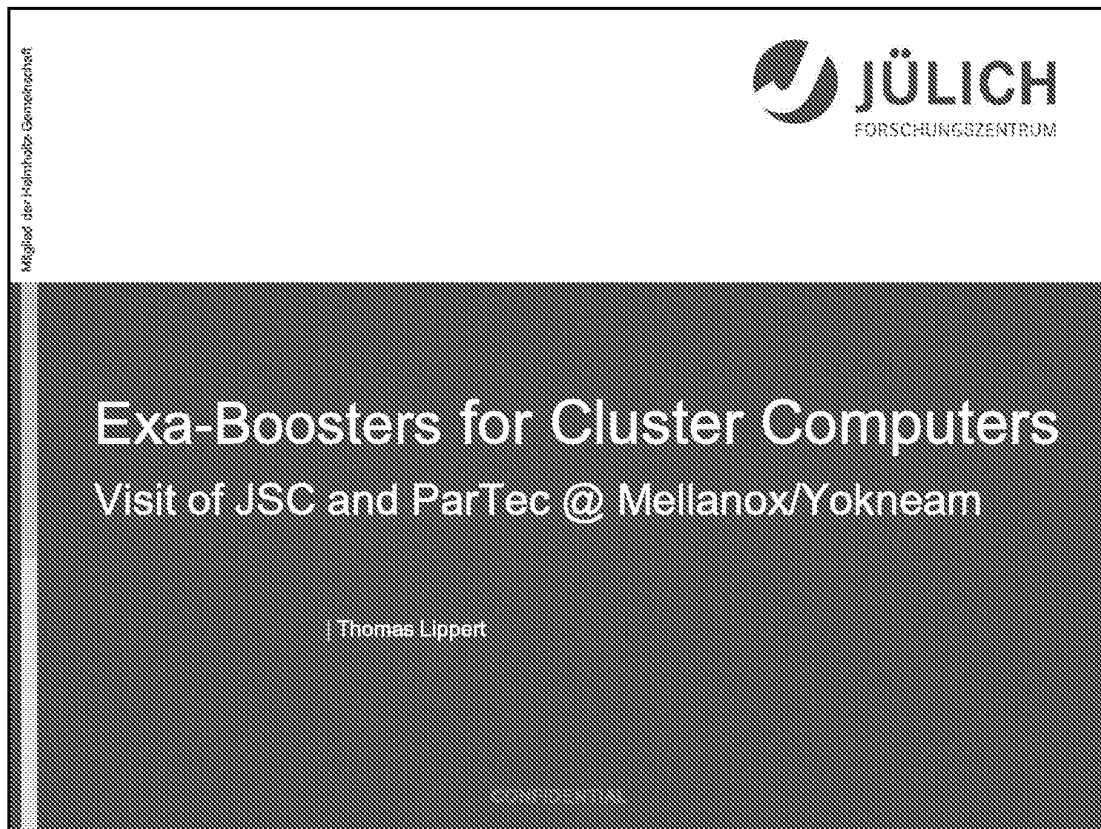



Exhibit A



Exhibit A



Exa-Cluster-Lab @ FZJ

Mission
Have a large impact on the development and realization of a sustained roadmap leading towards Exascale super-computers


Starting Point
JuRoPA Cluster technology (Hardware/Software)

Emphasis
General purpose, Novel concepts, Exascale performance, scalability and resilience

Partners
FZJ, Intel, ParTec

12.07.2010 CONFIDENTIAL 3

Exhibit A



Exa-Cluster-Lab: Staff


The lab is located in Jülich in order to benefit from experienced end-users and scientists of JSC and GRS

- 3 FTE's INTEL (Braunschweig and former Brühl (Pallas))
- 3 FTE's ParTec (located at Jülich)
- 3 FTE's JSC (newly created positions)
- To be extended to 15 staff

Operation has started 1st of July

12.07.2010 CONFIDENTIAL 4

Exhibit A




Exa-Cluster-Lab: 3 Initial Projects

- I. Open Exascale Software Stack - OES**
(FZJ, INTEL, PARTEC)
- II. Exa-Cluster Simulator**
(FZJ, INTEL)
replaced by
Exa-cluster Experimentation Platform - ECEP
(FZJ, INTEL, PARTEC)
- III. Scalable Exascale Tools - SET**
(FZJ, INTEL, PARTEC)

12.07.2010 CONFIDENTIAL 5

Exhibit A




WP1: OES

- ✦ OES: analysis and extension of existing software stacks and management layers for cluster supercomputers like JuRoPa for next generation supercomputer clusters
- ✦ Main focus will be set on the current limiting factors, as scalability, reliability, resiliency and performance.
- ✦ Starting with current open source ParaStation V5 cluster software, extensions of the software stack are planned
- ✦ Software developed will be open source under an open source license which is approved by the open source initiative (<http://www.opensource.org>).
- ✦ ParaStation V5 will be licensed without license fees or royalties to the lab, Intel and FZJ

12.07.2010 CONFIDENTIAL 6

Exhibit A




WP2: ECEP

- » Development of a Cluster Simulator removed in favor of an “ExaCluster Experimentation platform” using “Knights Ferry” devices
- » A Multi PCIX board will allow for testing the concept of a booster for clusters
- » ECEP will be the first step towards a future Knights Corner system
 - » Requires QPI license → ECL (!)
 - » Requires special network at least 10 x faster than today
 - » → EU funded development project?

12.07.2010 CONFIDENTIAL 7

Exhibit A



WP3: SET

- ✦ Beyond the system-level software Exascale require fundamentally new approaches on the application level.
- ✦ Existing programming tools will neither be capable to handle the multiple layers of parallelism Exascale-systems will exhibit nor application scalability.
- ✦ SEC will analyze and address the limitations of current Petascale software tools for tracing, monitoring and performance analysis and enable them for Exascale.
- ✦ Software developed shall be released as open source under an open source license which is approved by the open source initiative (<http://www.opensource.org>).

12.07.2010 CONFIDENTIAL 8

Exhibit A

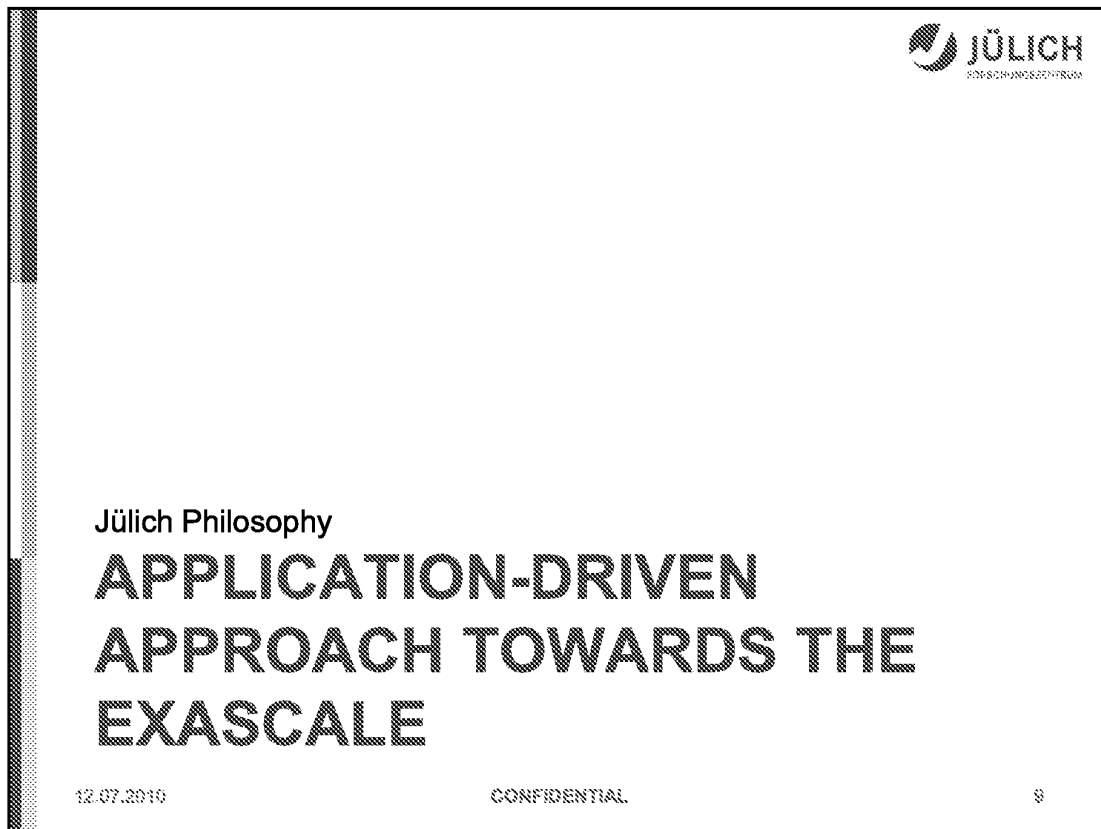



Exhibit A



The Jülich Dualistic Concept

- ※ 2004: Constellation systems found unable to scale
- ※ Portfolio of applications can be (very roughly) divided in two parts:
 - ※ Highly scalable codes, sparse-matrix vector like or dominated
 - ※ Highly complex codes, adaptive grids or coordinate based, all-to-all or more intricate communication patterns, large memory, less scalable
- ※ JSC decided to adapt hardware roadmap to this situation

12.07.2010 CONFIDENTIAL 19

Exhibit A

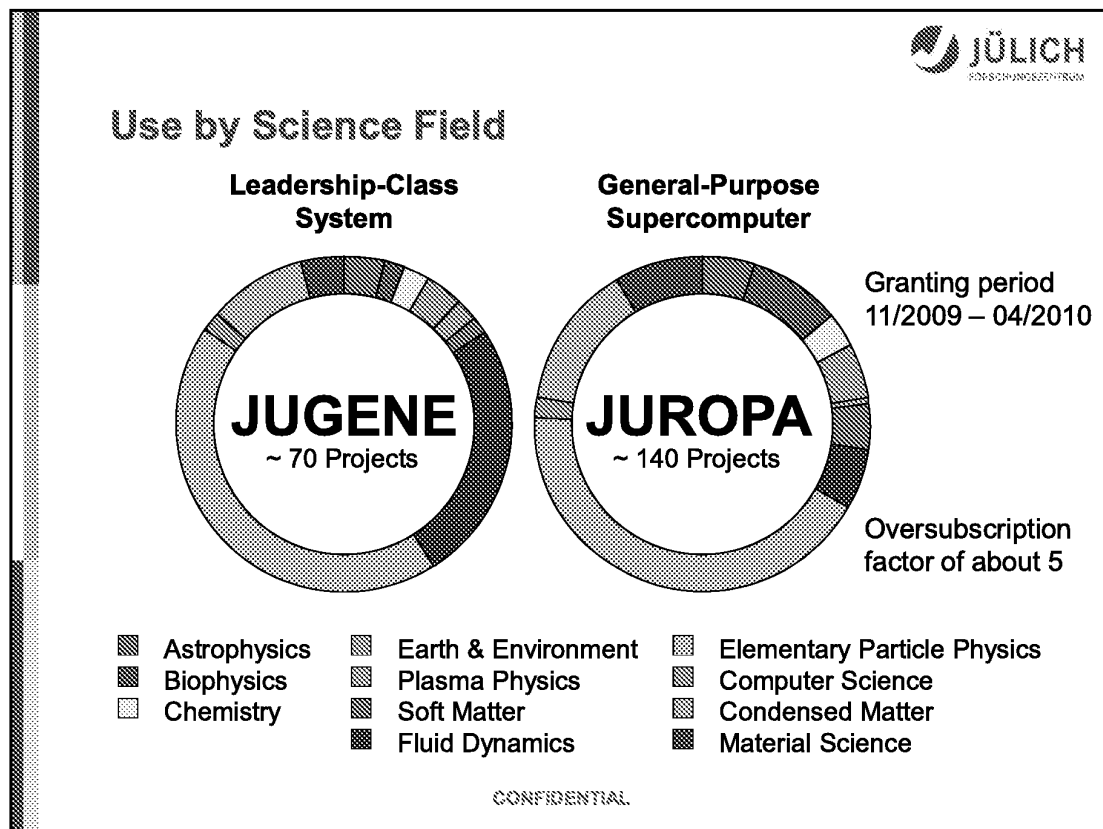


Exhibit A

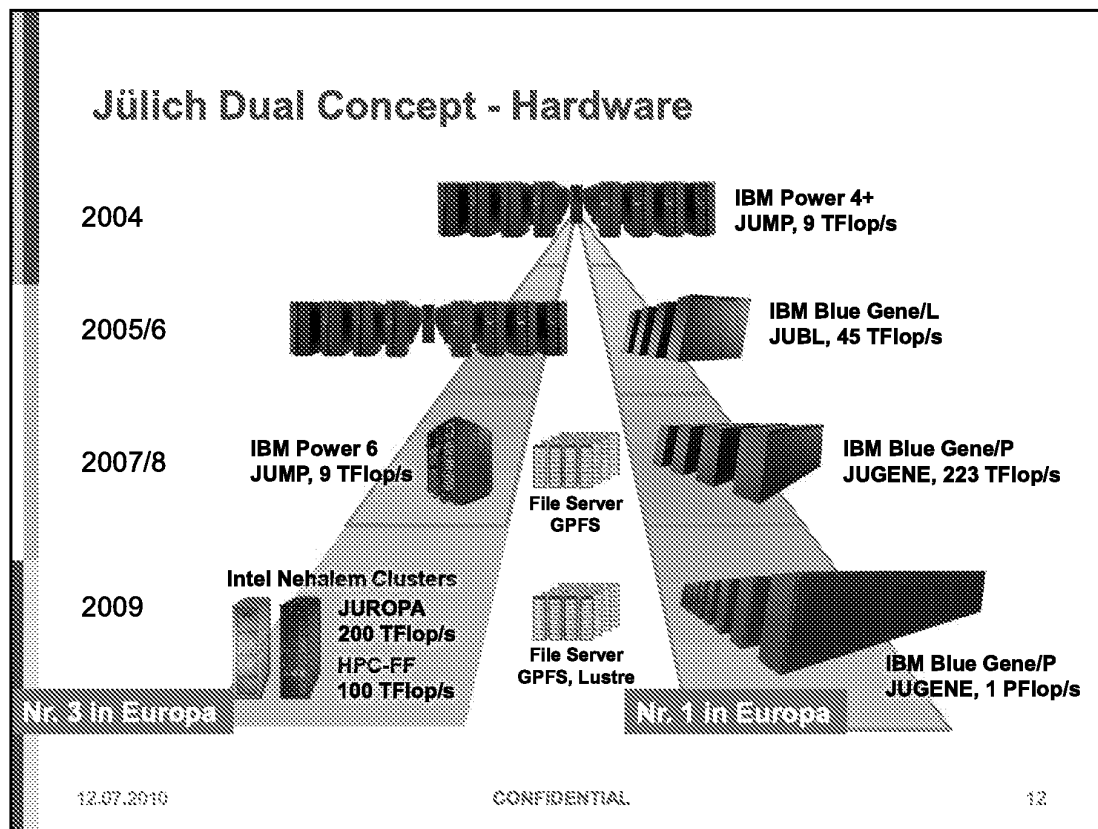



Exhibit A



A Closer Look on Application Codes

- ⌘ There is no puristic highly scalable code
- ⌘ There is no strictly complex code
- ⌘ → Each code has highly scalable and (less-scalable) complex elements
- ⌘ → There is a continuum between both extremes
- ⌘ Interestingly, many less scalable elements of a code do not require high scalability but instead large memory
- ⌘ All-to-all communication elements have a high advantage on smaller parallelism
- ⌘ Can we adapt the hardware architecture of future systems to take benefit from this situation?

12.07.2010 CONFIDENTIAL 13

Exhibit A

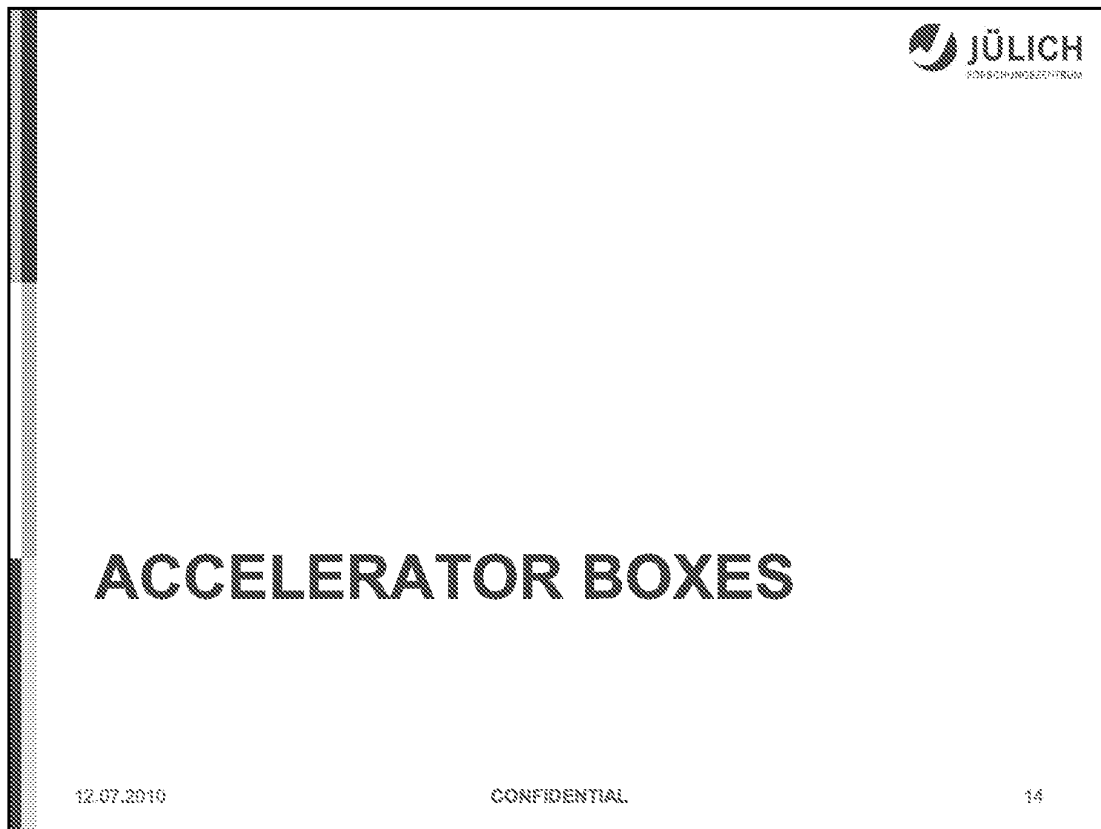



Exhibit A



Exhibit A




Rationale

- ◊ **Can the next generation cluster computers compete with proprietary solutions like Blue Gene or Cray?**
 - ※ Blue Gene /P → /Q gives factor 20 in compute speed at the same energy envelope and costs in 4 years
 - ※ Cray is more dependent on processor development
- ◊ **Standard processor speed will increase by about a factor of 4 to at most 8 in 4 years...**
 - ※ → Clusters need to utilize accelerators
 - ※ But: Current accelerators are not parallelized on the node-level
- ◊ **Integrated processors expected in 2015...**

12.07.2010 CONFIDENTIAL 16

Exhibit A




Rationale

- ◊ **Clusters going Exaflop/s will require virtualization elements in order to guarantee resilience and reliability.**
 - ※ → Virtualization software layer
- ◊ **Local accelerators are static elements**
 - ※ → Have to tolerate over/under subscription
 - ※ → No fault tolerance if accelerator fails
 - ※ → very difficult to virtualize
 - ※ But: can utilize high local bandwidth
 - ※ Allow a simple view of the system
 - ※ Work well for farming or master-slave parallelization

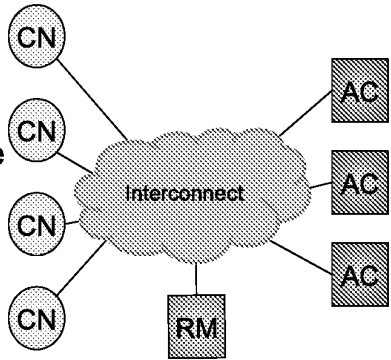
12.07.2010 CONFIDENTIAL 17

Exhibit A

 JÜLICH
FORSCHUNGSZENTRUM

Booster (inspired by discussion in February)

- ✧ Coupling of accelerators (AC) to compute nodes (CN) is flexible
- ✧ Sharing of accelerators between compute nodes possible
- ✧ Virtualization on cluster level possible
- ✧ AC-to-CN assignment goes by resource manager (RM):
 - ✧ Static (at job start)
 - ✧ Dynamic (at runtime)
- ✧ Cluster-Accelerator communication proceeds through network protocol
- ✧ Intra-Accelerator communication requires new solution




12.07.2010 CONFIDENTIAL 18

Exhibit A



Exhibit A




Advantages

- Dynamic and static AC-to-CN assignment
- Virtualization of cluster is not hampered
- Exploit accelerator parallelism
- Accelerator allocation follows application needs
- Fault tolerance in case of accelerator failure
- All compute nodes share same growth capacity
- Potential for O(10) PF in 2014

12.07.2010 CONFIDENTIAL 20

Exhibit A



CONS


- IB network extension required
- Specific fast network among accelerators required
- Specific boards to be developed
- QPI or HT will be essential between accelerators
- Need to develop enabling middleware layer and math libraries as well as compiler technology and programming models

12.07.2010 CONFIDENTIAL 21

Exhibit A



Exhibit A




Exascale Software Initiatives and Potential Funding

- **IESP:** International Exascale Software Project
 - Led by DOE (ANL/ORNL – Beckmann/Dongarra)
- **EESI:** European Exascale Software Initiative (EDF France)
- **F77:** ICT Objective „HPC Platzforms with Exascale Performance“ – K. Glinos in preparation
- **PRACE** Second implementation phase: scaling applications
- **FET Flagship Initiative Supercomputing**
 - (Technology beyond 2020 – BSC, INRIA, JSC)
- **G8:** Interdisciplinary Program on Application Software towards Exascale Computing for Global Scale Issues

12.07.2010 CONFIDENTIAL 23

Exhibit A




ICT-2011.9.13

The FP7 ICT call 7 will present the “Objective ICT-2011.9.13 Exascale computing, software and simulation”. This is the first objective in FP7 dedicated to Exascale technology. It will be supported through probably three integrated projects (IP) with a volume of M€ 8 each.

12.07.2010 CONFIDENTIAL 24

Exhibit A




Exascale Call

- EC Exa-call expected in September
- 3 projects à 8 M€
- Need one leading Centre per project
- Need one leading development company
- Need very strong and innovative European technology component
- Need universities (mainly applications)

12.07.2010 CONFIDENTIAL 25

Exhibit A




What is intended

- **Proposals should address extreme parallelism with millions of cores and should involve algorithms, programming models, compilers, power consumption, etc.**
- **Each IP should comprise one or more SC centres, technology and system suppliers including vendors, industrial or academic centres to co-develop a small number of Exa-scaled application codes. 40 % application software, 60 % system development.**
- **Proposals may include international cooperation components. All software should be developed as open source. EC officers have confirmed that there can be a US company involved and that a few SMEs from one country might contribute. The IP should involve at least three of more countries from Europe.**

12.07.2010 CONFIDENTIAL 26

Exhibit A




Dynamical Exa-Cluster Computing (DECO)

- ※ **Supercomputer Centres**
 - ※ JSC (Leading), TUM (multi core) (tbc), BSC (compiling techn) (Spain)
- ※ **Accelerator**
 - ※ INTEL: new Intel Many Core Processors (X86 programming model) – Knights Ferry – Knights Corner (US)
- ※ **Network**
 - ※ Mellanox: Inter-cluster communication, cluster-to-accelerator communication (Israel)
 - ※ EXTOLL: 3d accelerator network (Germany)
 - ※ ParTec: Dynamical Exa-cluster OS (Bavaria)
- ※ **System**
 - ※ Eurotec (Italy)
- ※ **Applications**
 - ※ Lausanne: Human Brain Project (Switzerland)
 - ※ CERFACS: Fluid engineering (France)
 - ※ Leuven: Space Weather (Belgium)

12.07.2010 CONFIDENTIAL 27

Exhibit A



Mellanox


- ◊ Today
 - ※ HCA: CX2, PCI2.0 x 8 , 2e IB QDR, 40 GbitE or 10 GbitE., L 850 ns
 - ※ Switch IS4, QDR, L 100 ns
 - ※ Bridge
- ◊ Q4 2010
 - ※ HCA: CX3 PCIX 3.0 x 8
 - ※ FDR (14 Gbit) → 56 Gbit
 - ※ Switch 36 ports IB, ETH, FBC
- ◊ H2 2011
 - ※ HCA: Golan, PCIX 3.0 x 16
 - ※ 2 x EDR

12.07.2010

CONFIDENTIAL

28

Exhibit A




- ◊ H1/2012
 - ※ Golan, EDR, PCIX = 128 Gbit unidir
 - ※ Switch DER, DER=100 Gbit per direction

12.07.2010

CONFIDENTIAL

29

Exhibit A



Next Steps

- ✧ Meet Brüning (2.9.)
- ✧ Contact Intel...Intel Plans?
- ✧ Contact Eurotec until end of July

12.07.2010

CONFIDENTIAL

30